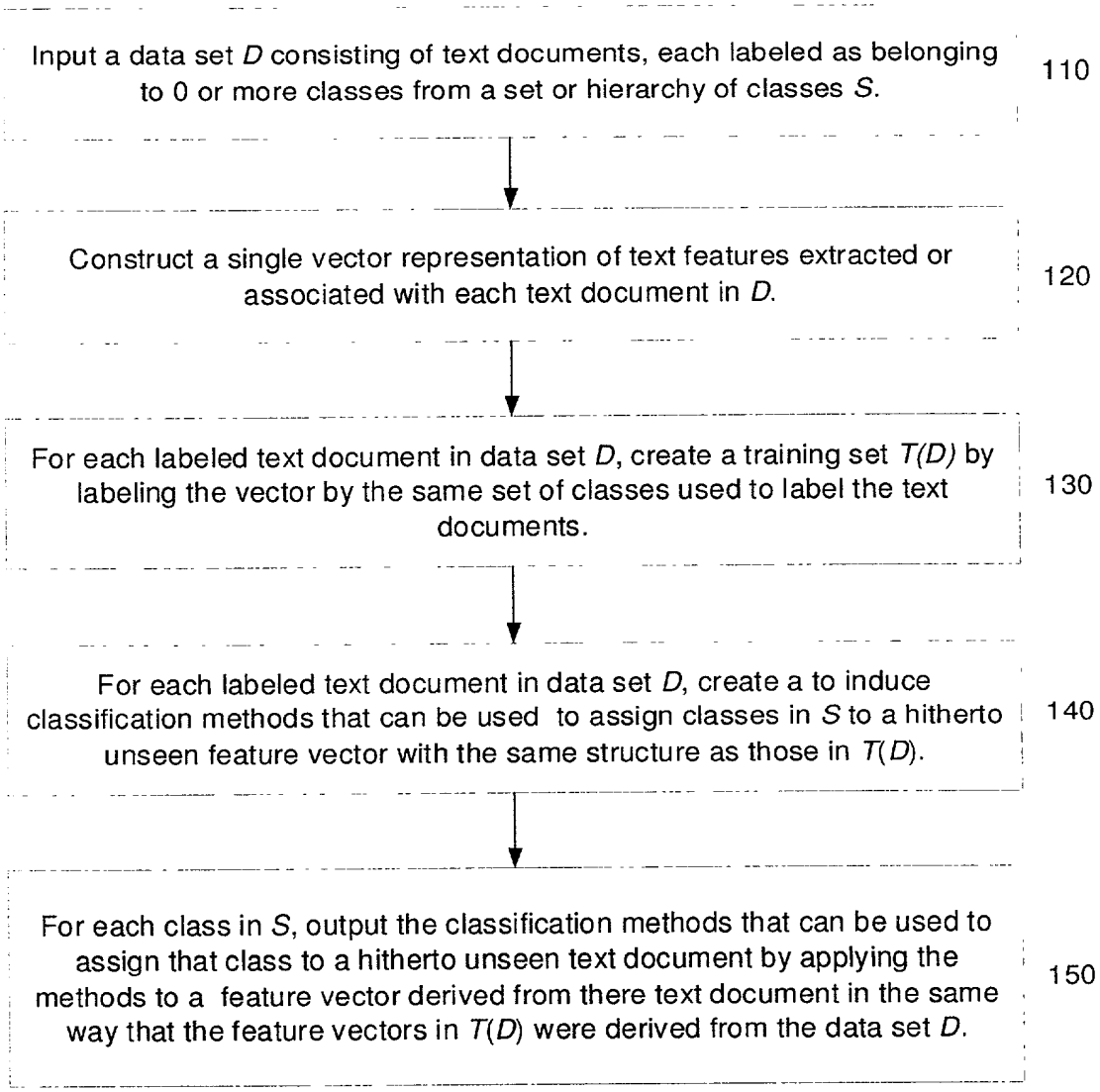


Prior Art

100

Figure 1

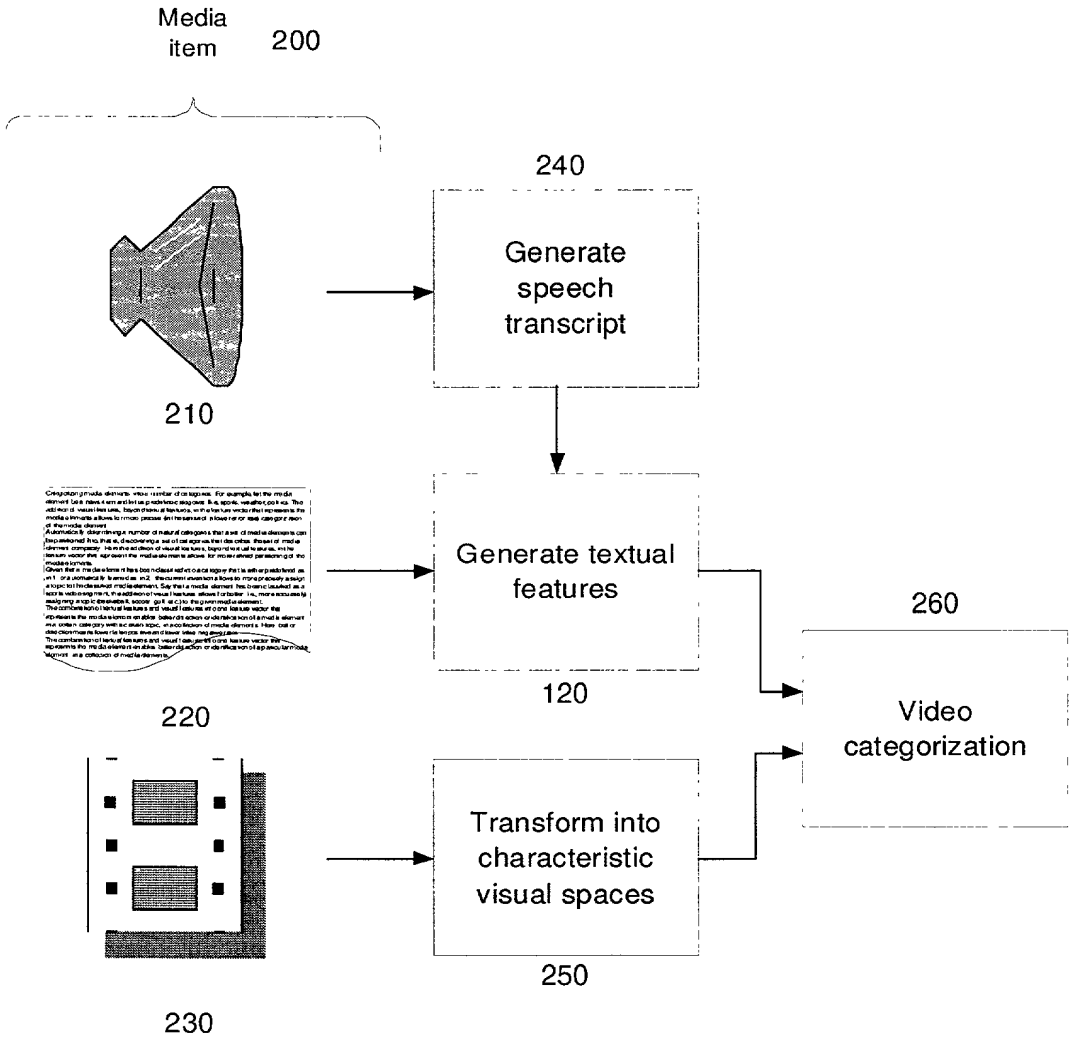

Media item
200
240
Generate speech transcript
210
Generate textual features
220
120
260
Video categorization
Transform into characteristic visual spaces
230
250

Figure 2

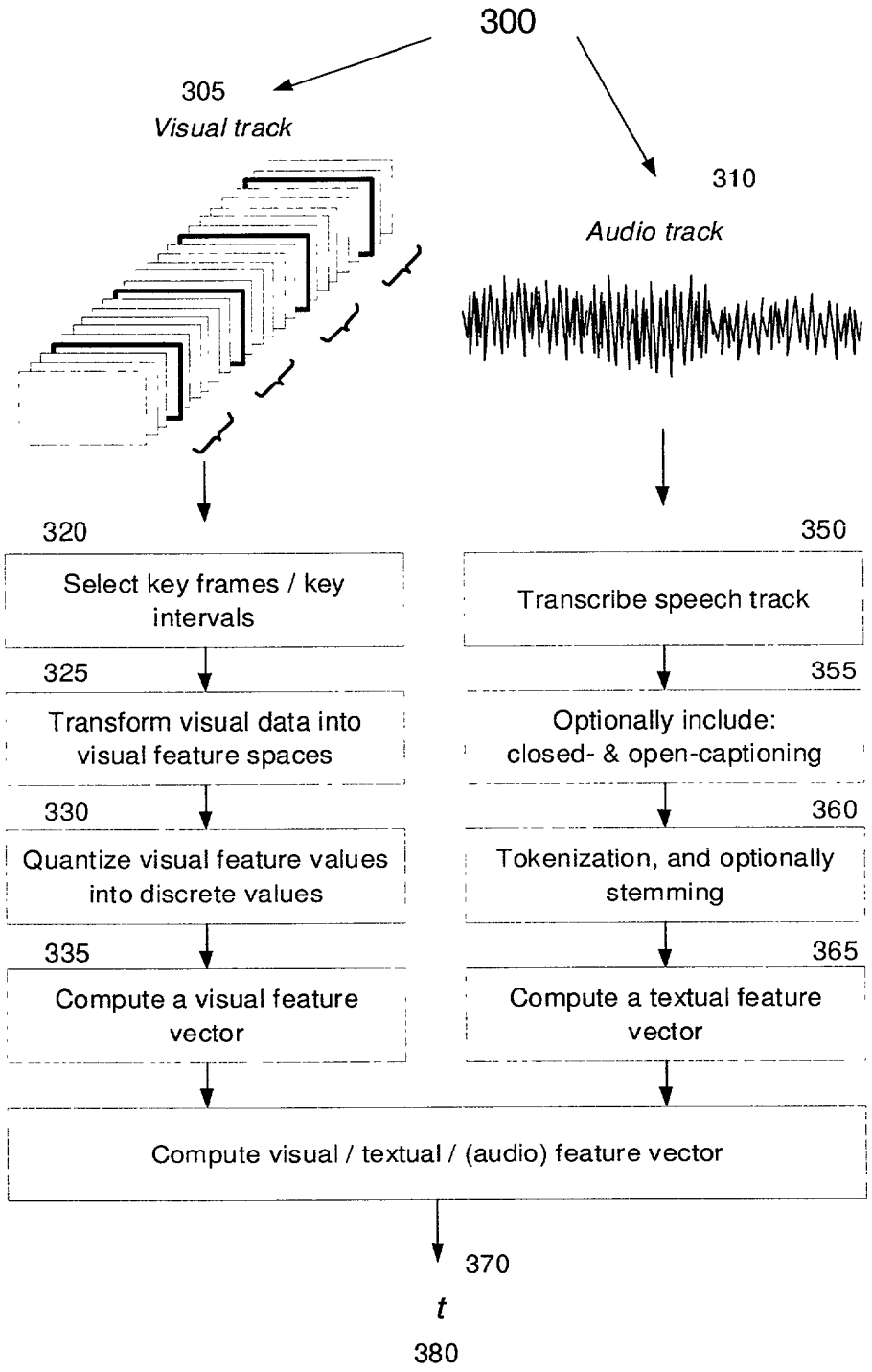

300
305
Visual track
310
Audio track
320
Select key frames / key intervals
325
Transform visual data into visual feature spaces
330
Quantize visual feature values into discrete values
335
Compute a visual feature vector
350
Transcribe speech track
355
Optionally include: closed- & open-captioning
360
Tokenization, and optionally stemming
365
Compute a textual feature vector
Compute visual / textual / (audio) feature vector
370
t
380

Figure 3

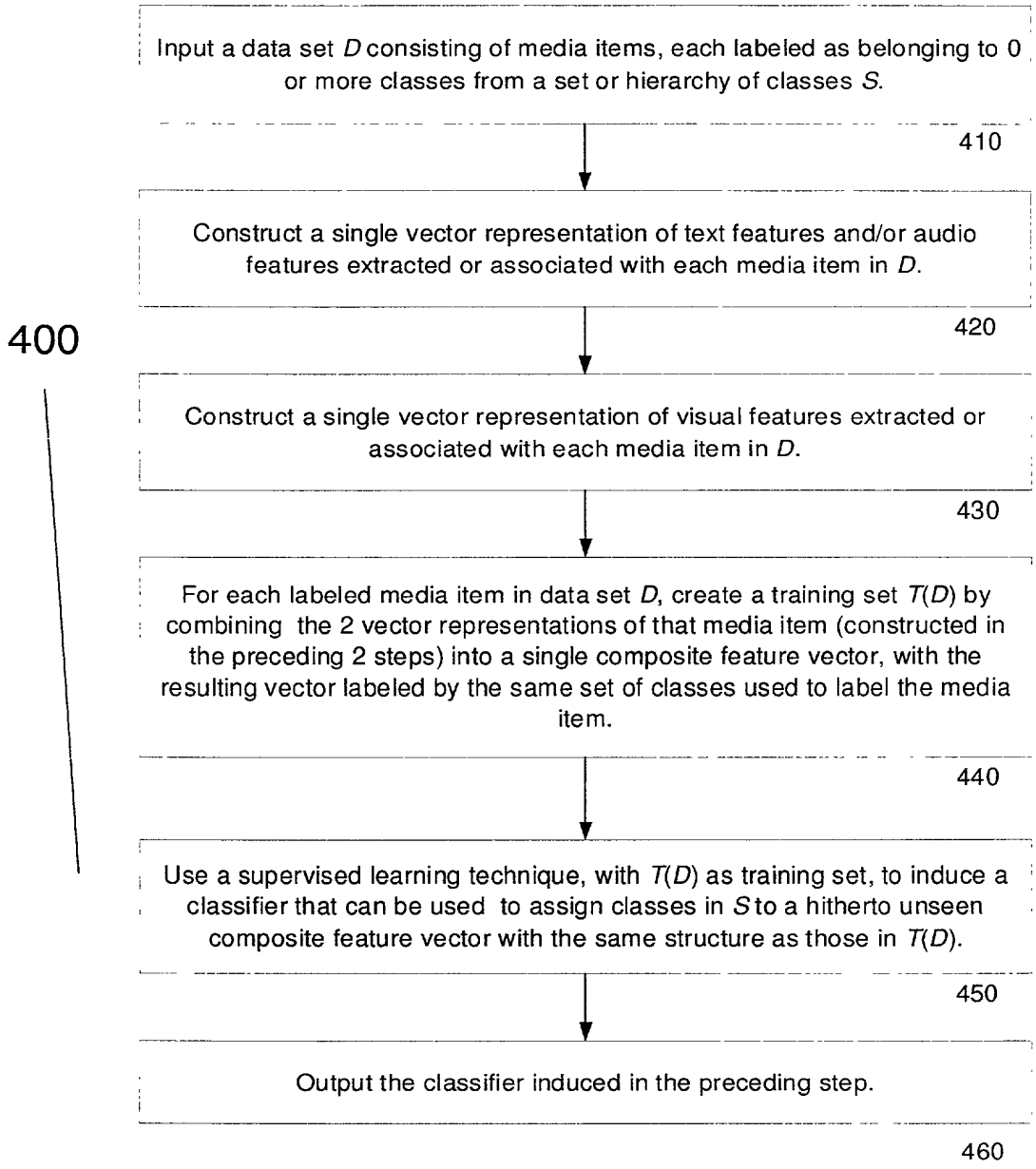

400
Input a data set D consisting of media items, each labeled as belonging to 0 or more classes from a set or hierarchy of classes S.
410
Construct a single vector representation of text features and/or audio features extracted or associated with each media item in D.
420
Construct a single vector representation of visual features extracted or associated with each media item in D.
430
For each labeled media item in data set D, create a training set T(D) by combining the 2 vector representations of that media item (constructed in the preceding 2 steps) into a single composite feature vector, with the resulting vector labeled by the same set of classes used to label the media item.
440
Use a supervised learning technique, with T(D) as training set, to induce a classifier that can be used to assign classes in S to a hitherto unseen composite feature vector with the same structure as those in T(D).
450
Output the classifier induced in the preceding step.
460

Figure 4

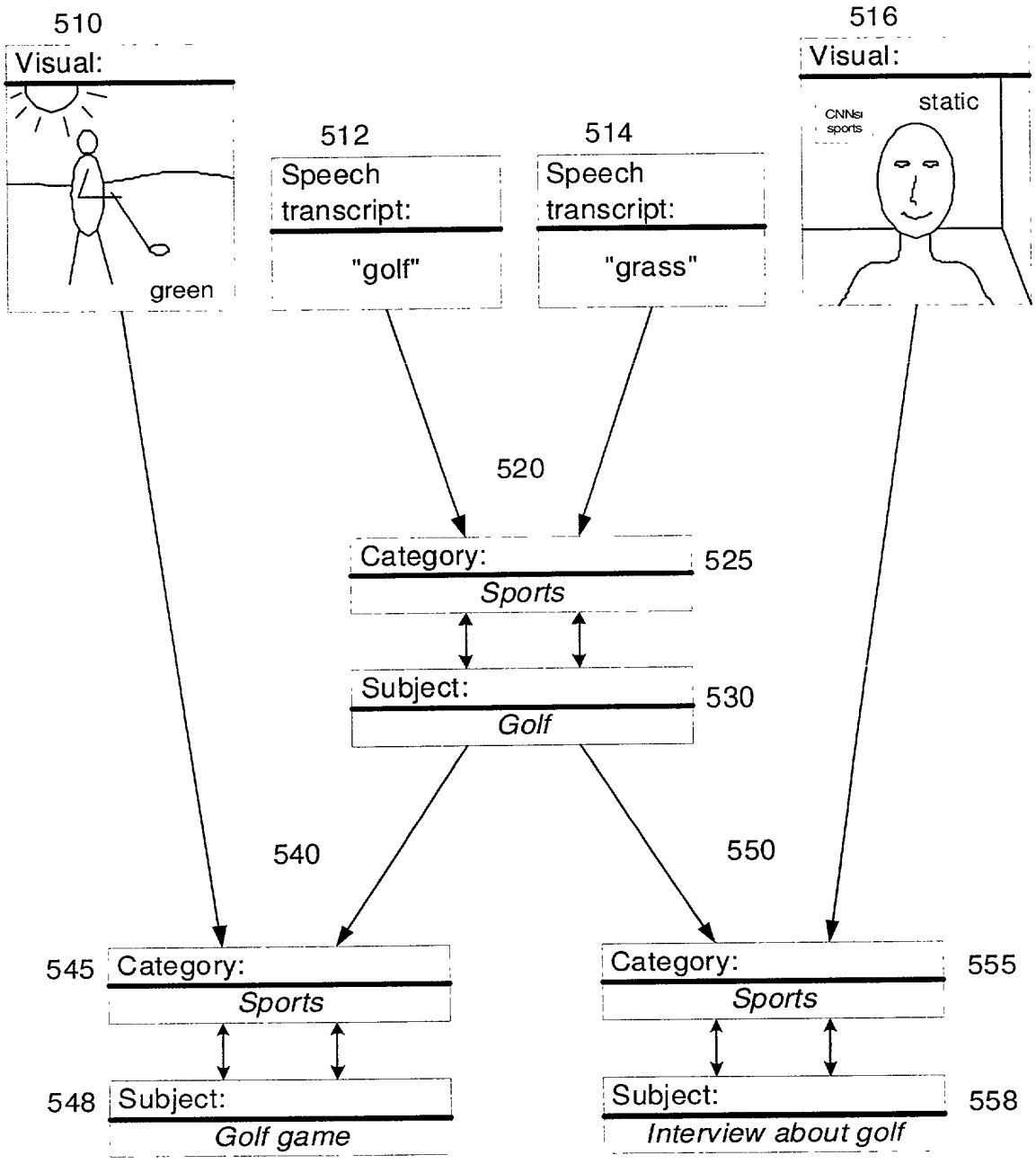
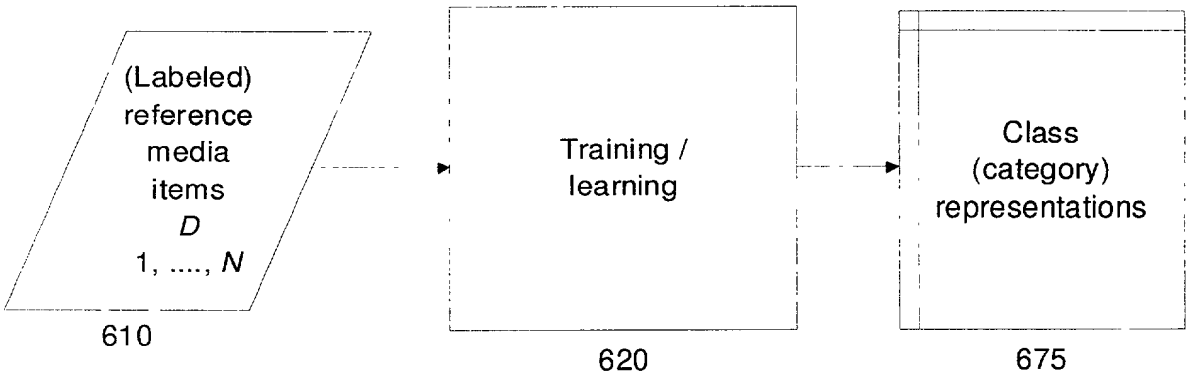
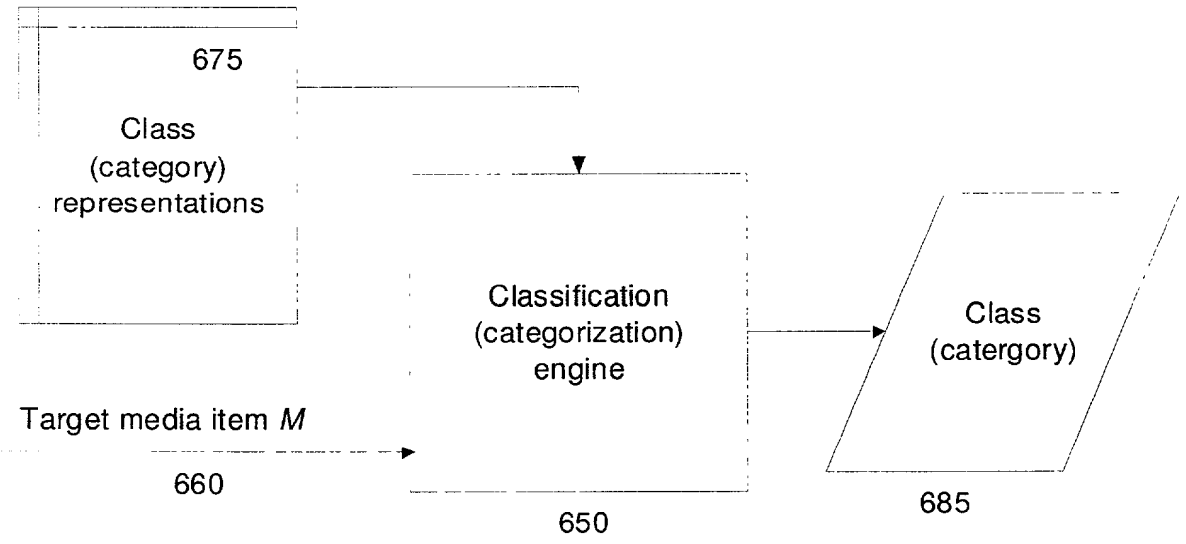

510
Visual:
green
512
Speech transcript:
"golf"
514
Speech transcript:
"grass"
516
Visual:
static
CNNsi sports
520
Category:
Sports
525
Subject:
Golf
530
540
550
545
Category:
Sports
548
Subject:
Golf game
Category:
Sports
555
Subject:
Interview about golf
558

Figure 5



Training / learning phase **610**

A



Classification phase **615**

B

Figure 6

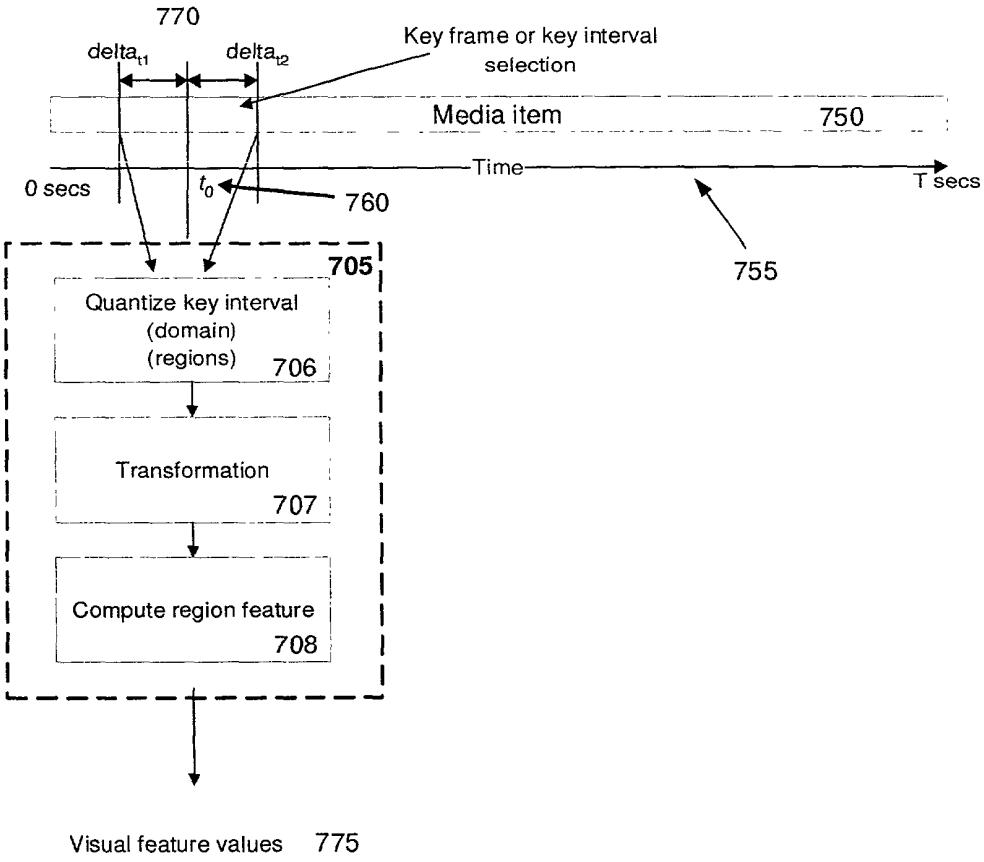

770
delta$_{t1}$
delta$_{t2}$
Key frame or key interval selection
Media item
750
Time
0 secs
t_0
760
T secs
755
705
Quantize key interval (domain) (regions)
706
Transformation
707
Compute region feature
708
Visual feature values
775

Figure 7A

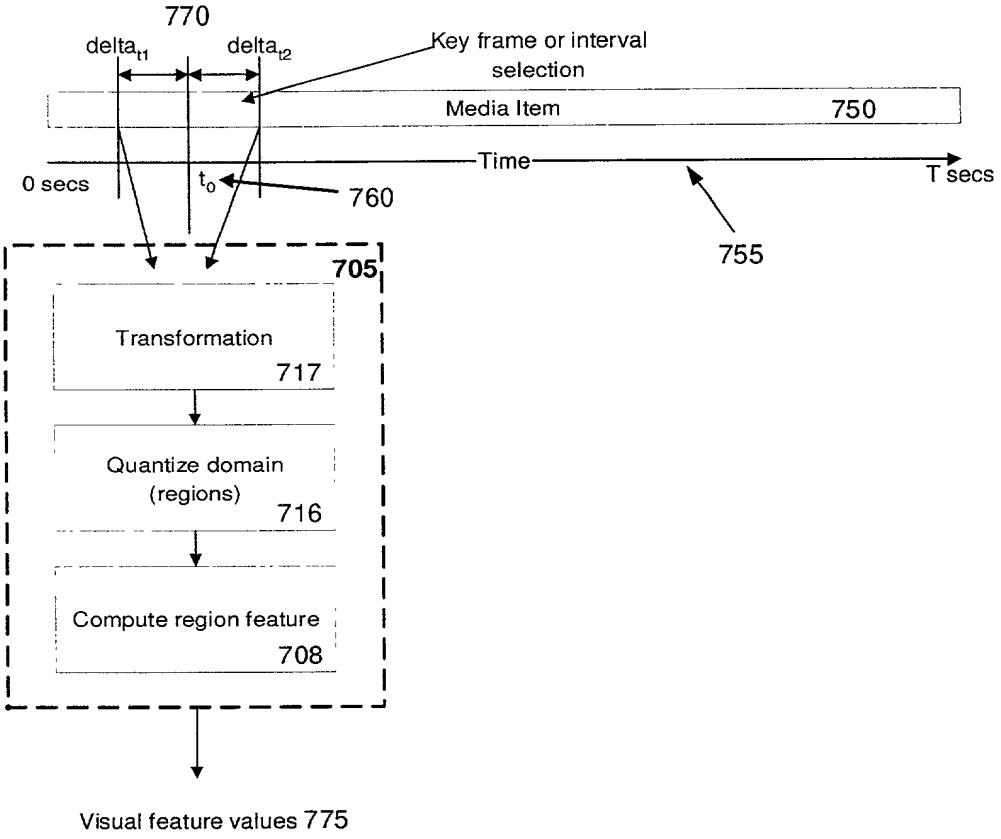

770
delta$_{t1}$
delta$_{t2}$
Key frame or interval selection
Media Item
750
Time
0 secs
t_0
760
T secs
755
705
Transformation
717
Quantize domain (regions)
716
Compute region feature
708
Visual feature values 775

Figure 7B

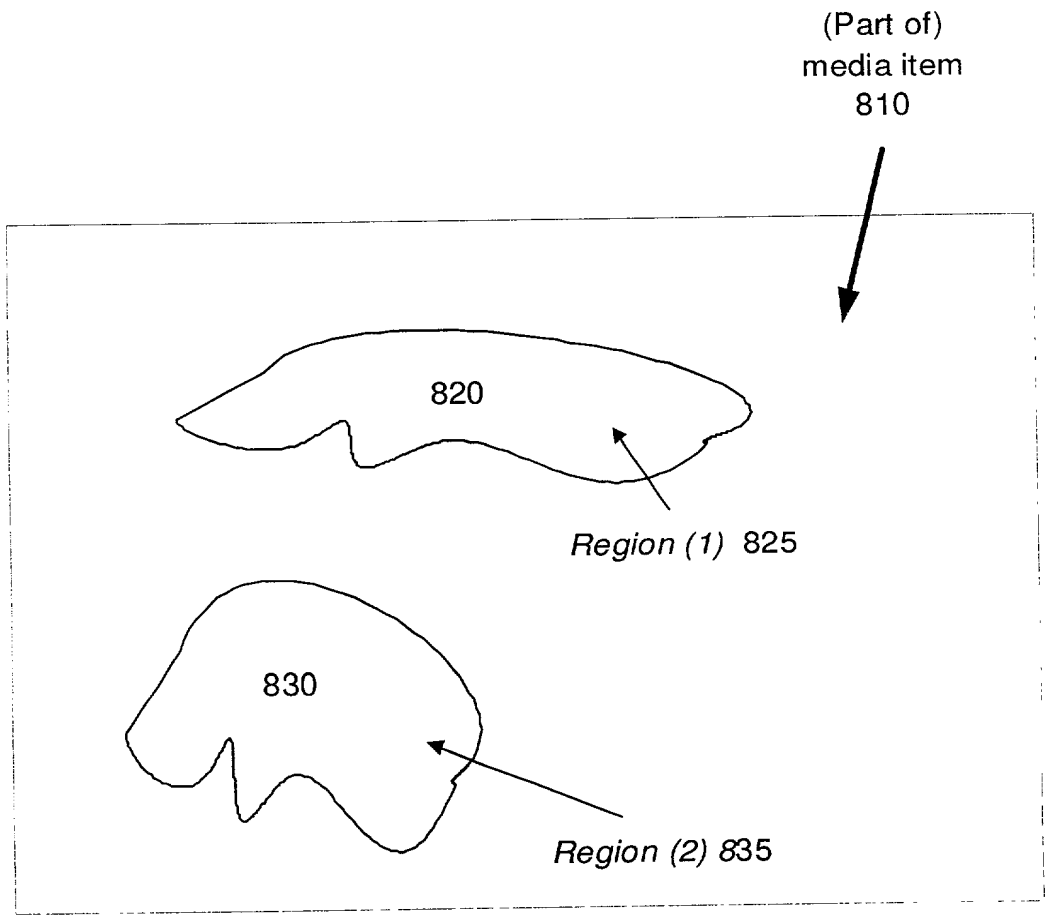


Figure 8

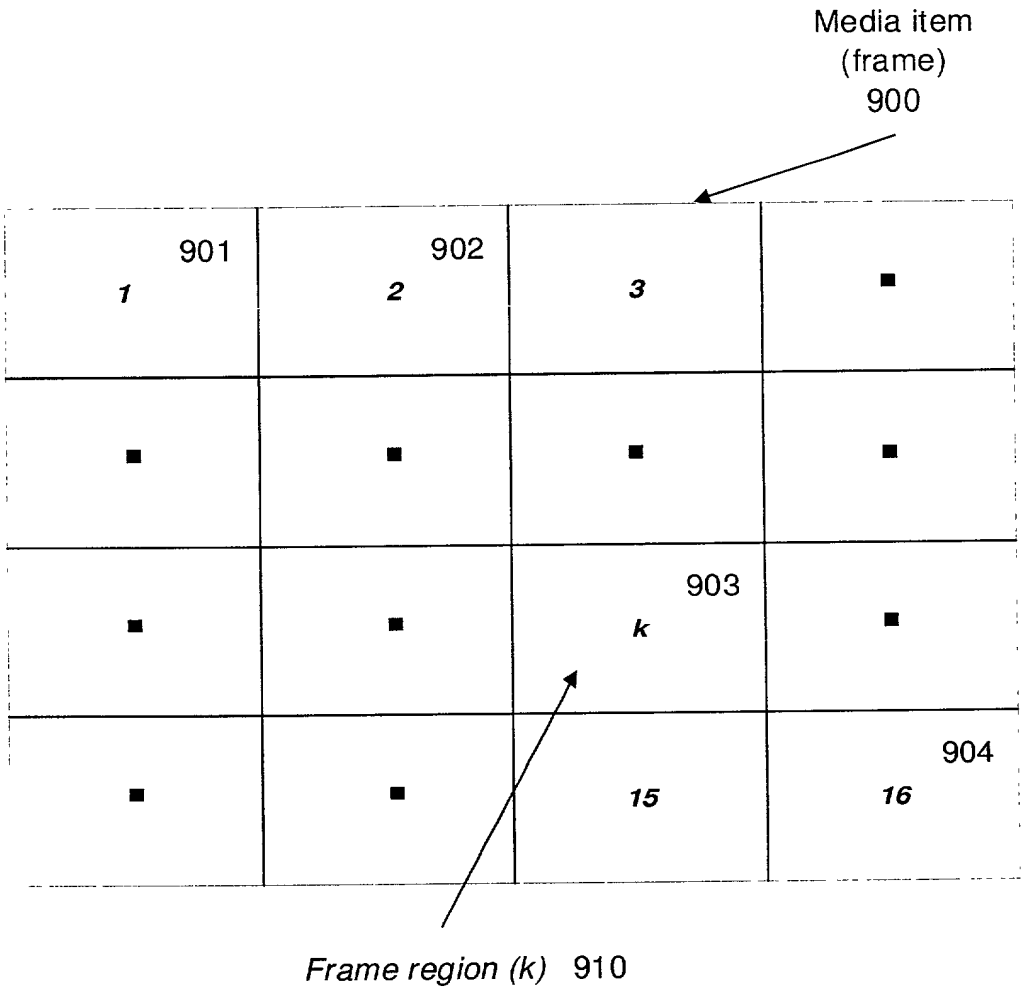

Media item
(frame)
900
901
1
902
2
3
903
k
904
15
16
Frame region (k) 910

Figure 9

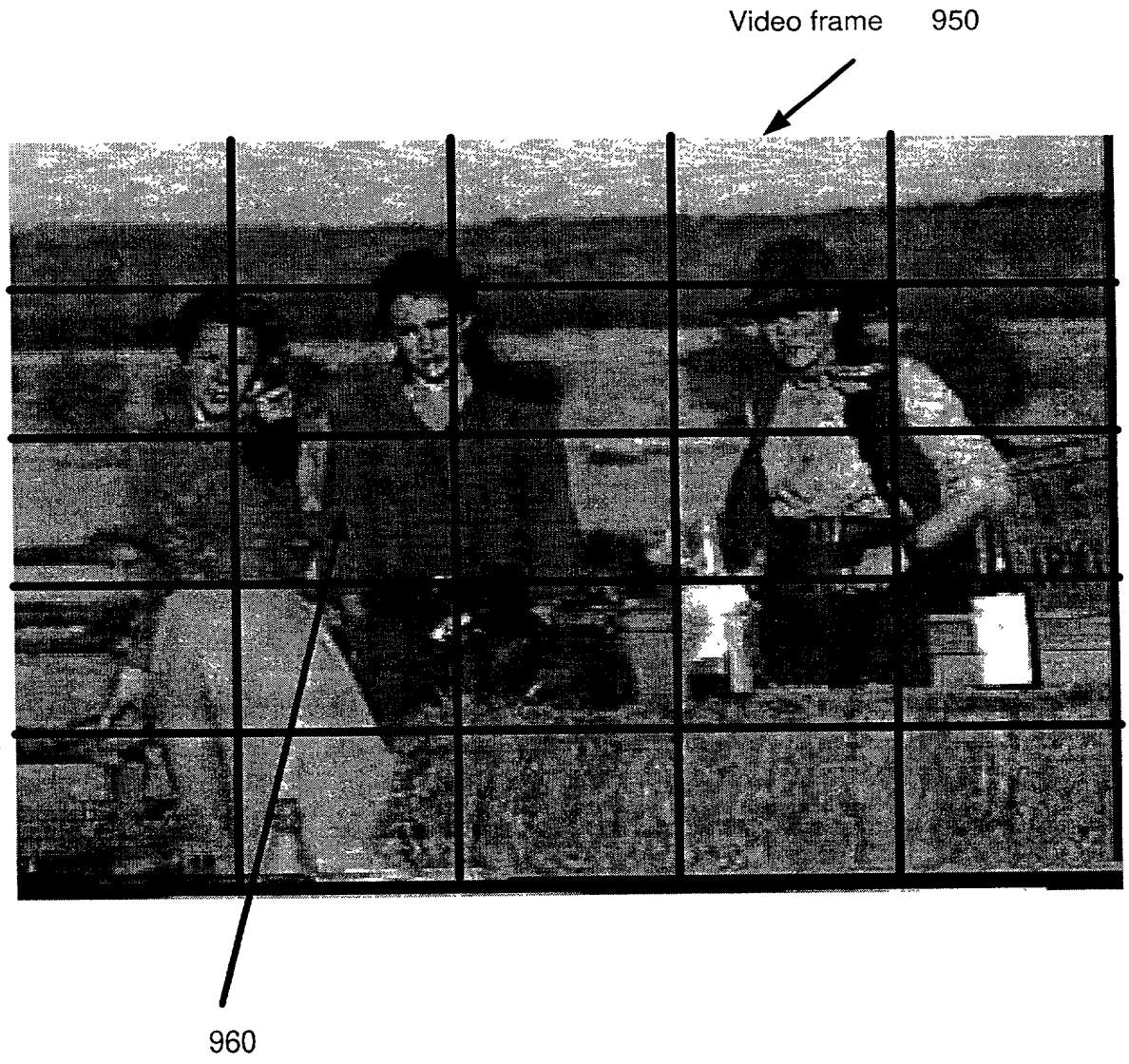


Figure 9A

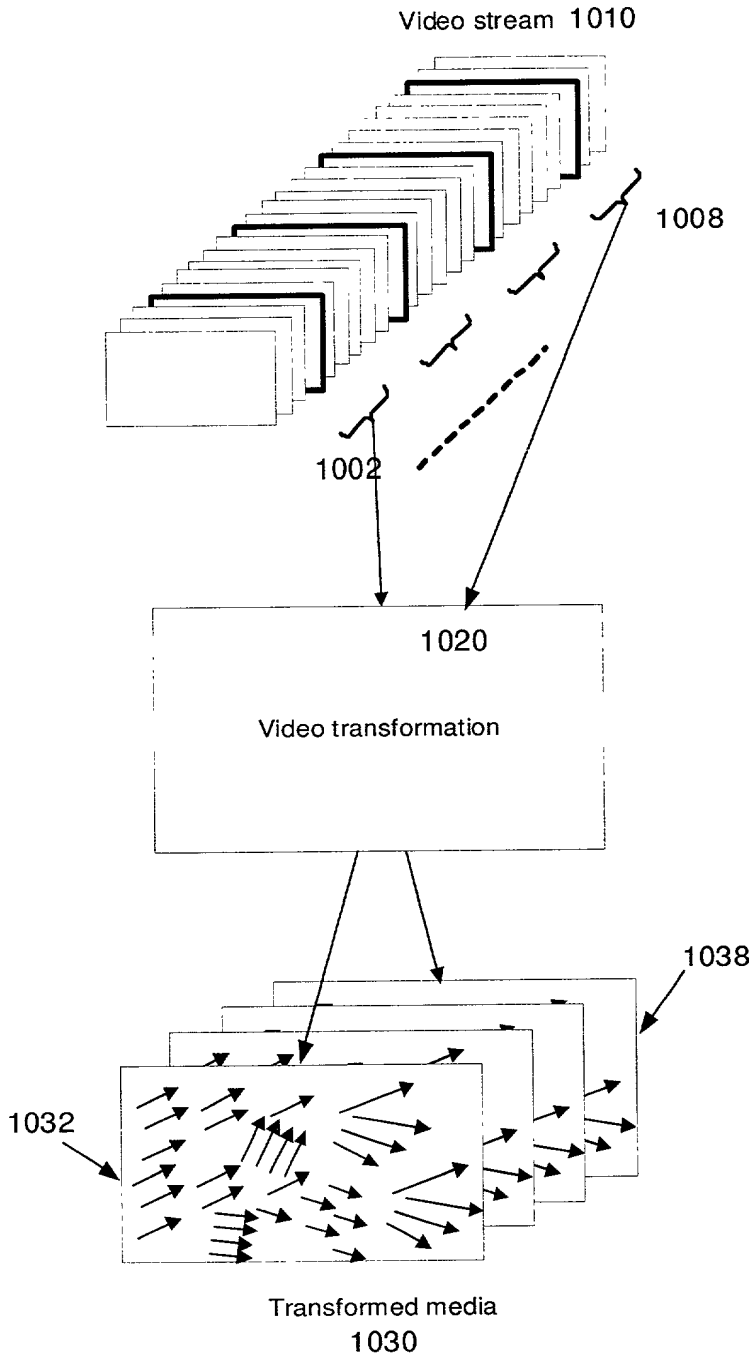

Video stream 1010
1008
1002
1020
Video transformation
1038
1032
Transformed media
1030

Figure 10

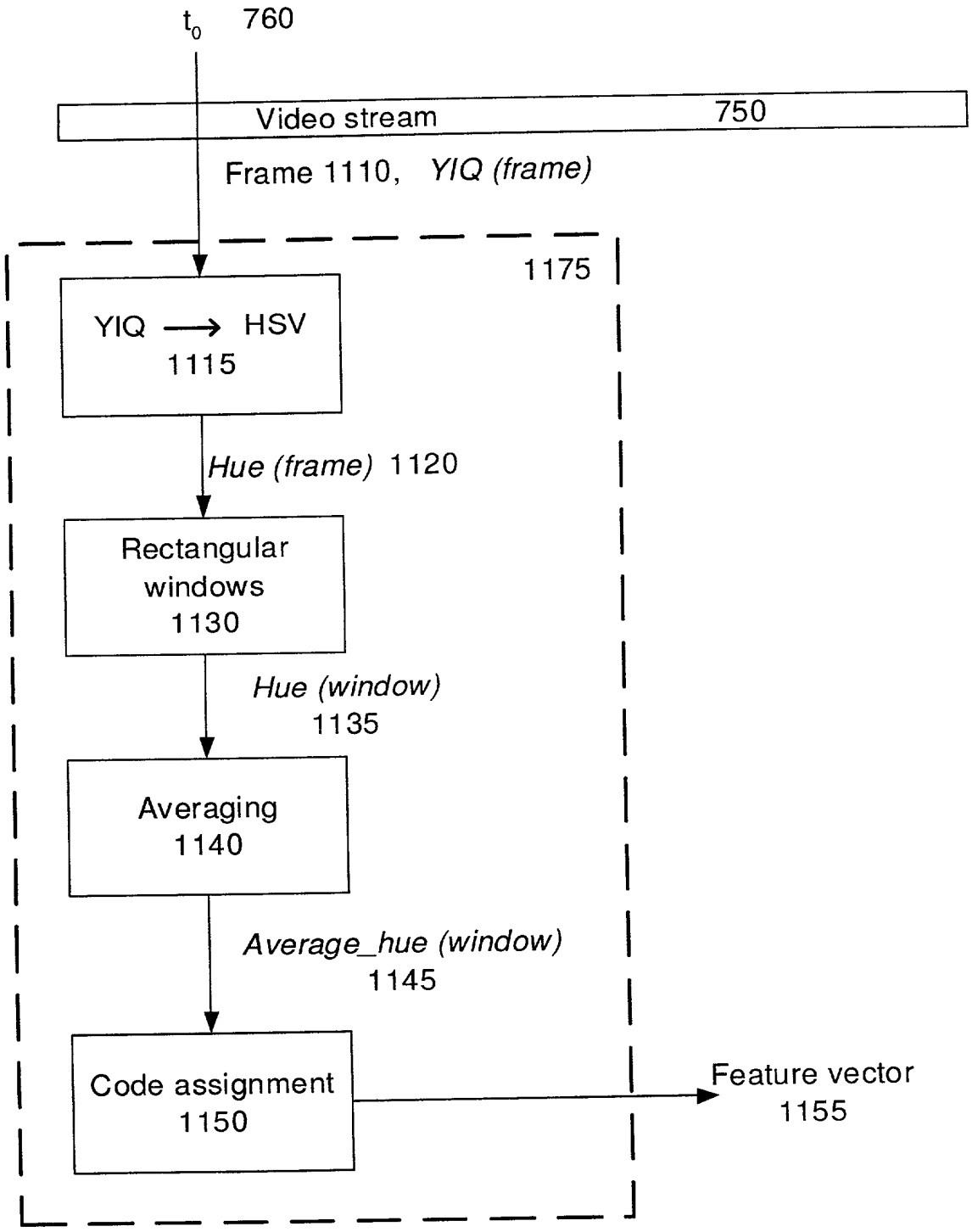
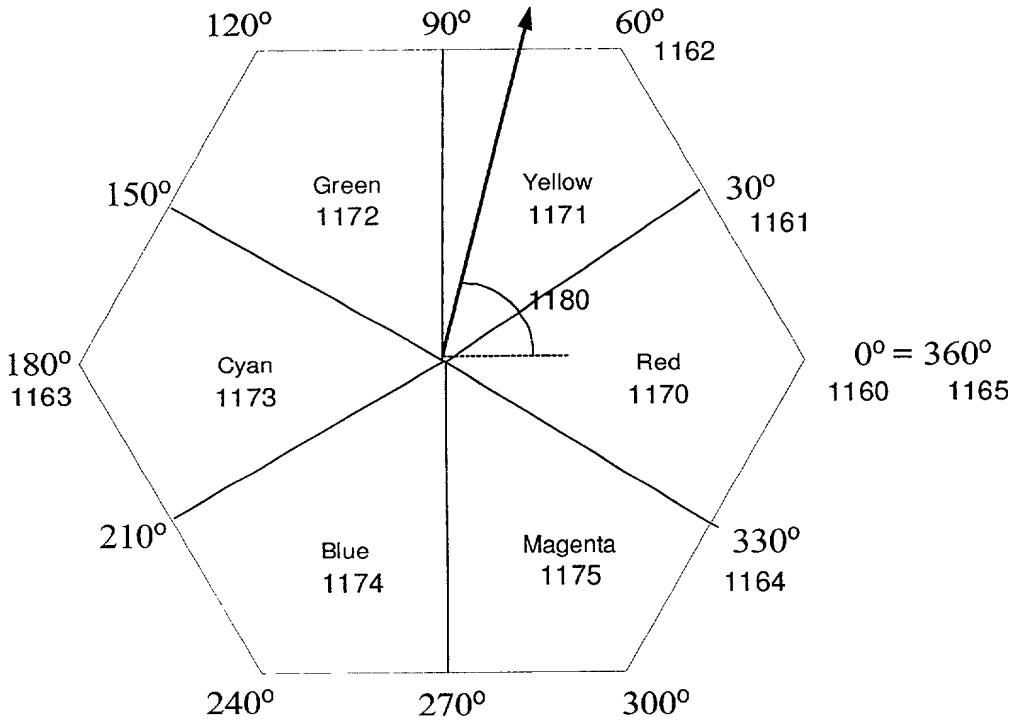

t_0 760
Video stream
750
Frame 1110, YIQ (frame)
1175
YIQ → HSV
1115
Hue (frame) 1120
Rectangular windows
1130
Hue (window)
1135
Averaging
1140
Average_hue (window)
1145
Code assignment
1150
Feature vector
1155

Figure 11A



Hue quantization (1180)

Color range		*Code*	*Color*	
330 - 30	degrees	0	Red	(**1170**)
30 - 90	degrees	1	Yellow	(**1171**)
90 - 150	degrees	2	Green	(**1172**)
150 - 210	degrees	3	Cyan	(**1173**)
210 - 270	degrees	4	Blue	(**1174**)
270 - 330	degrees	5	Magenta	(**1175**)

Figure 11B

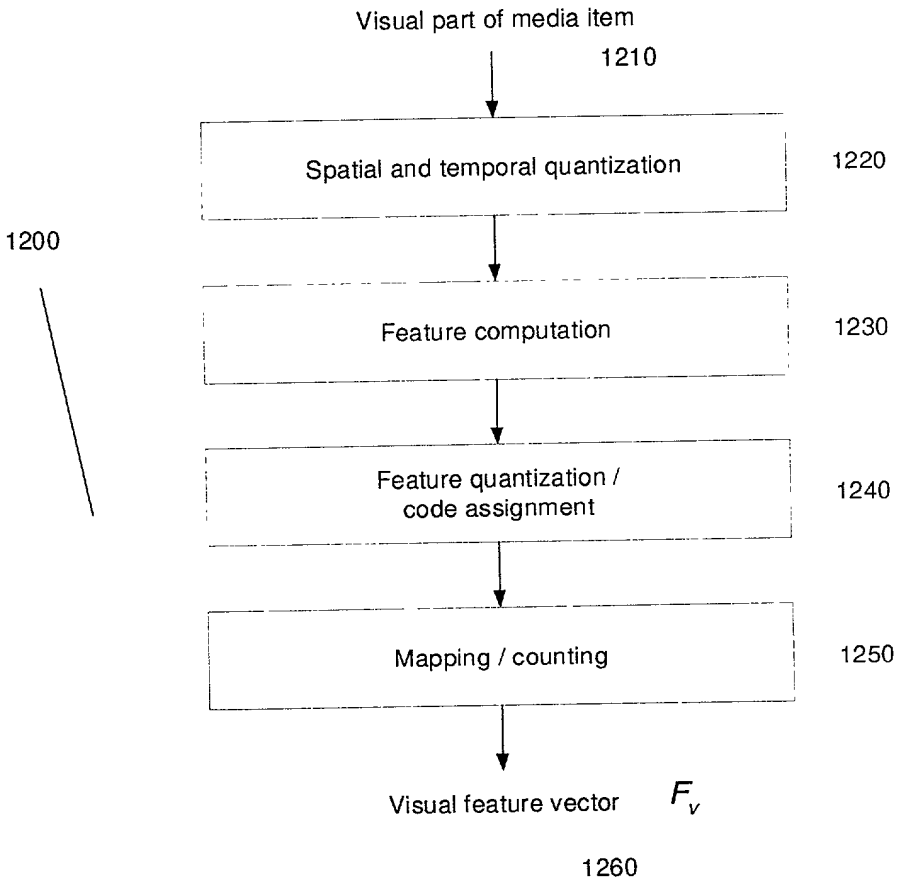

Visual part of media item
1210
Spatial and temporal quantization
1220
1200
Feature computation
1230
Feature quantization /
code assignment
1240
Mapping / counting
1250
Visual feature vector
F_v
1260

Figure 12

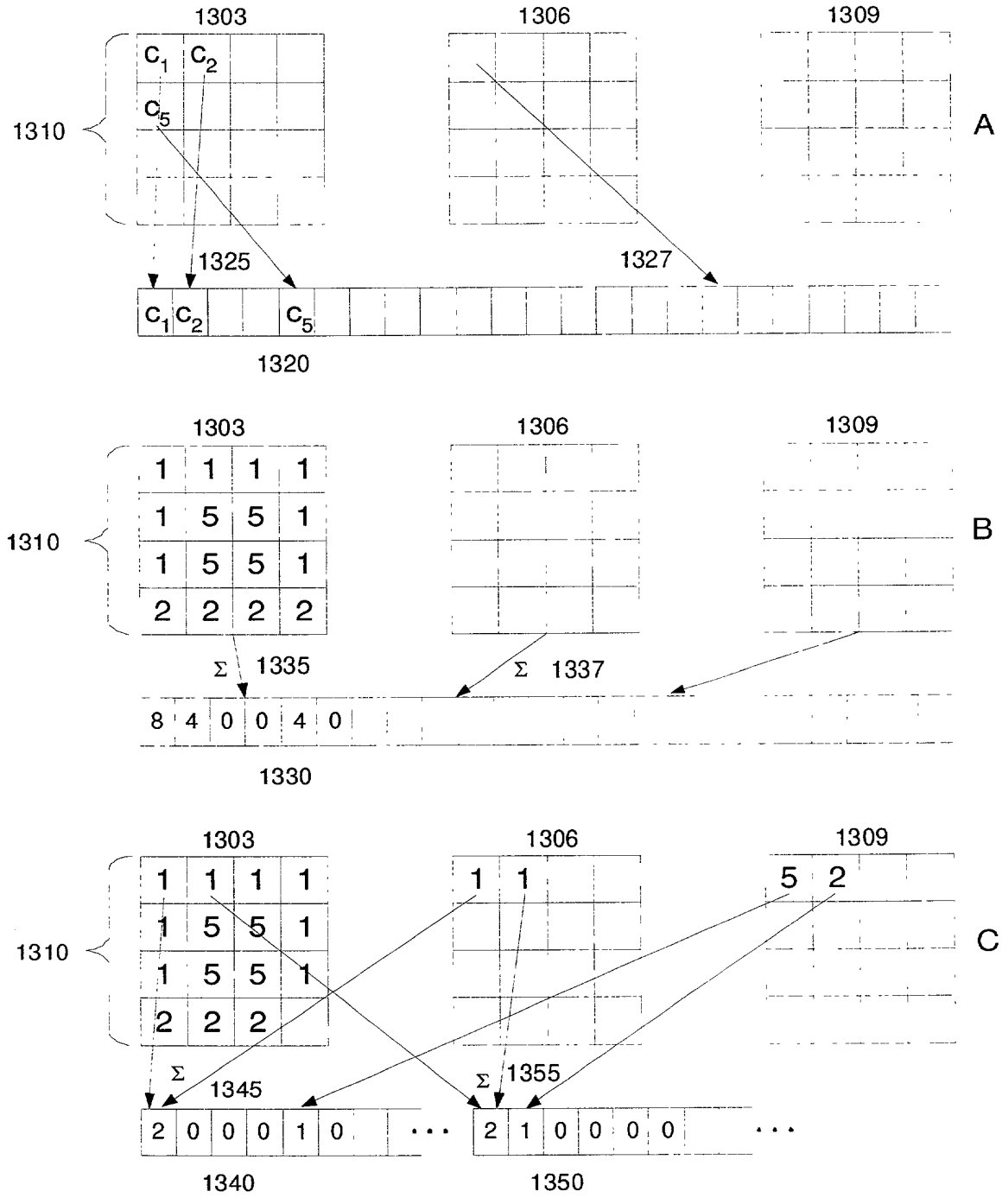

1303
1306
1309
1310
A
C_1 C_2
C_5
1325
1327
C_1 C_2 C_5
1320
1303
1306
1309
1310
B
1 1 1 1
1 5 5 1
1 5 5 1
2 2 2 2
Σ 1335
Σ 1337
8 4 0 0 4 0
1330
1303
1306
1309
1310
C
1 1 1 1
1 5 5 1
1 5 5 1
2 2 2
1 1
5 2
Σ
1345
Σ 1355
2 0 0 0 1 0 ...
2 1 0 0 0 0 ...
1340
1350

Figure 13

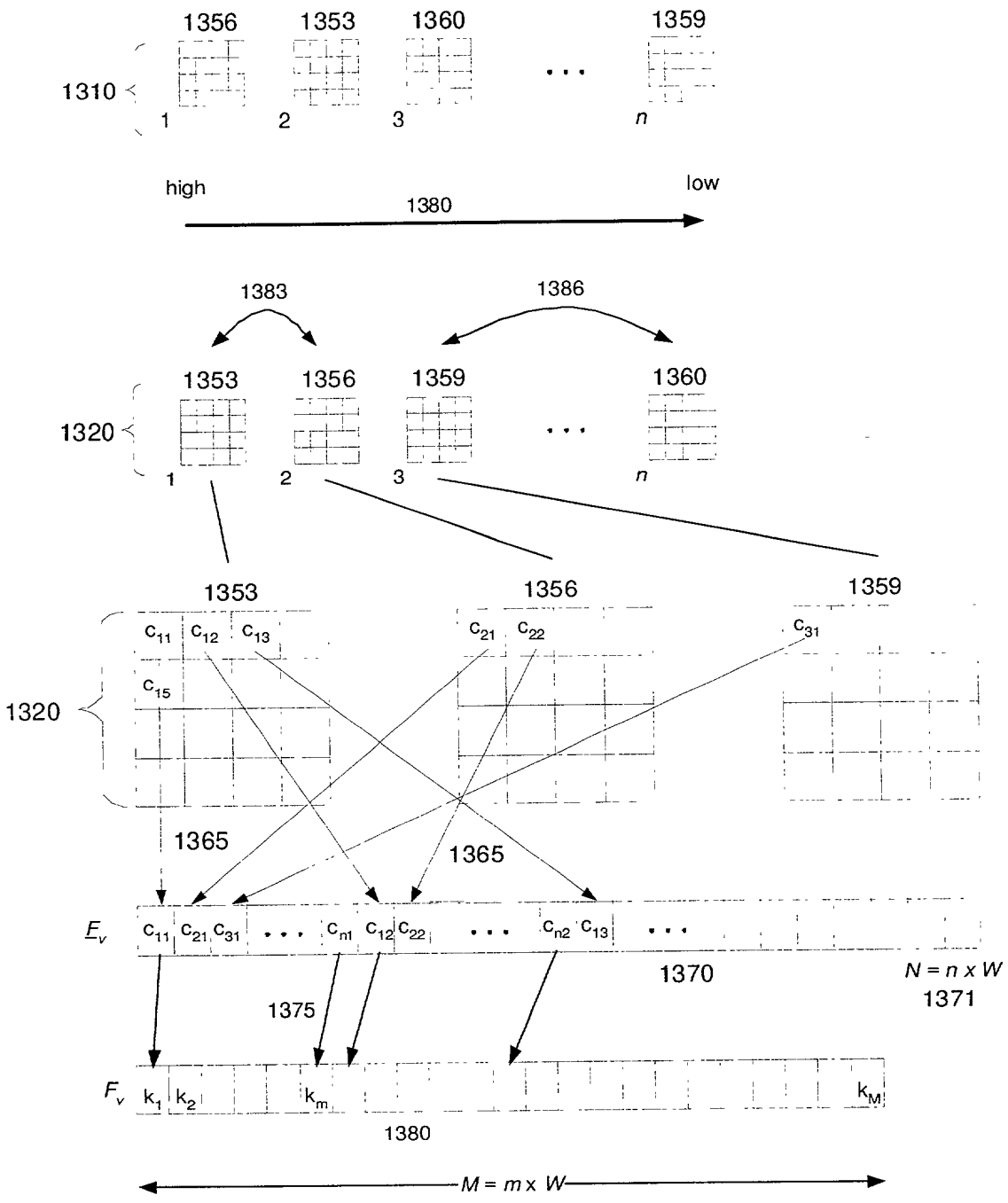
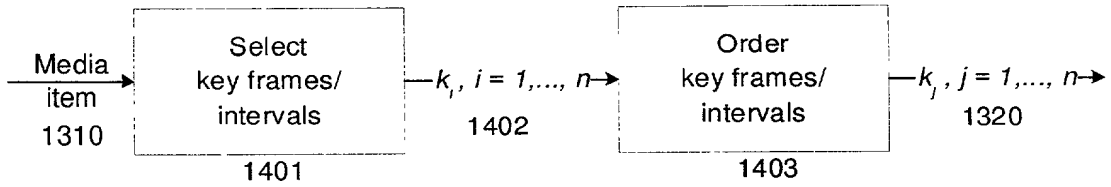
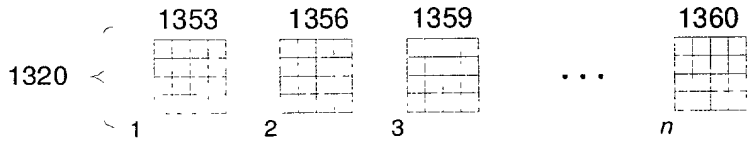

1356
1353
1360
1359
1310
1
2
3
n
high
low
1380
1383
1386
1353
1356
1359
1360
1320
1
2
3
n
1353
1356
1359
1320
c_{11} c_{12} c_{13}
c_{15}
c_{21} c_{22}
c_{31}
1365
1365
F_v
c_{11} c_{21} c_{31} ... c_{n1} c_{12} c_{22} ... c_{n2} c_{13} ...
1370
$N = n \times W$
1371
1375
F_v
k_1 k_2 k_m k_M
1380
$M = m \times W$

Figure 13D


Media
item
1310
Select
key frames/
intervals
1401
k_i, i = 1,..., n
1402
Order
key frames/
intervals
1403
k_j, j = 1,..., n
1320


1320
1353
1356
1359
1360
1
2
3
n

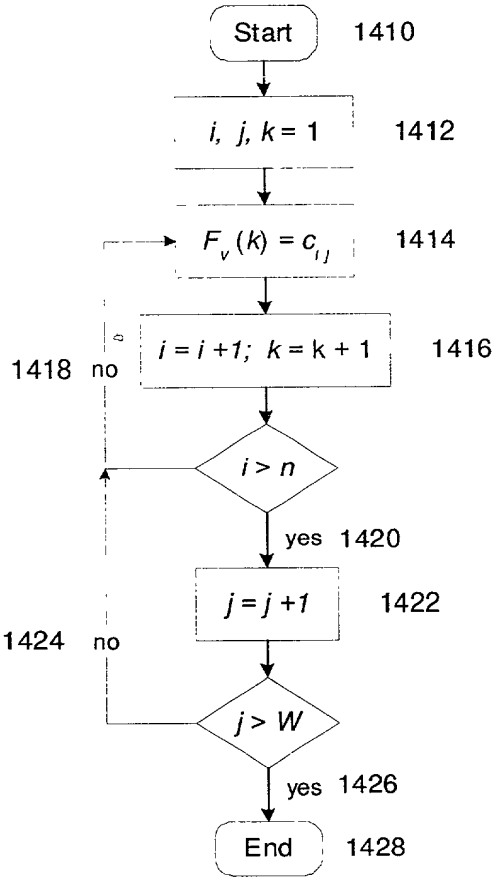

Start
1410
i, j, k = 1
1412
$F_v(k) = c_{ij}$
1414
i = i +1; k = k + 1
1416
1418
no
i > n
yes
1420
j = j +1
1422
1424
no
j > W
yes
1426
End
1428

Figure 14A

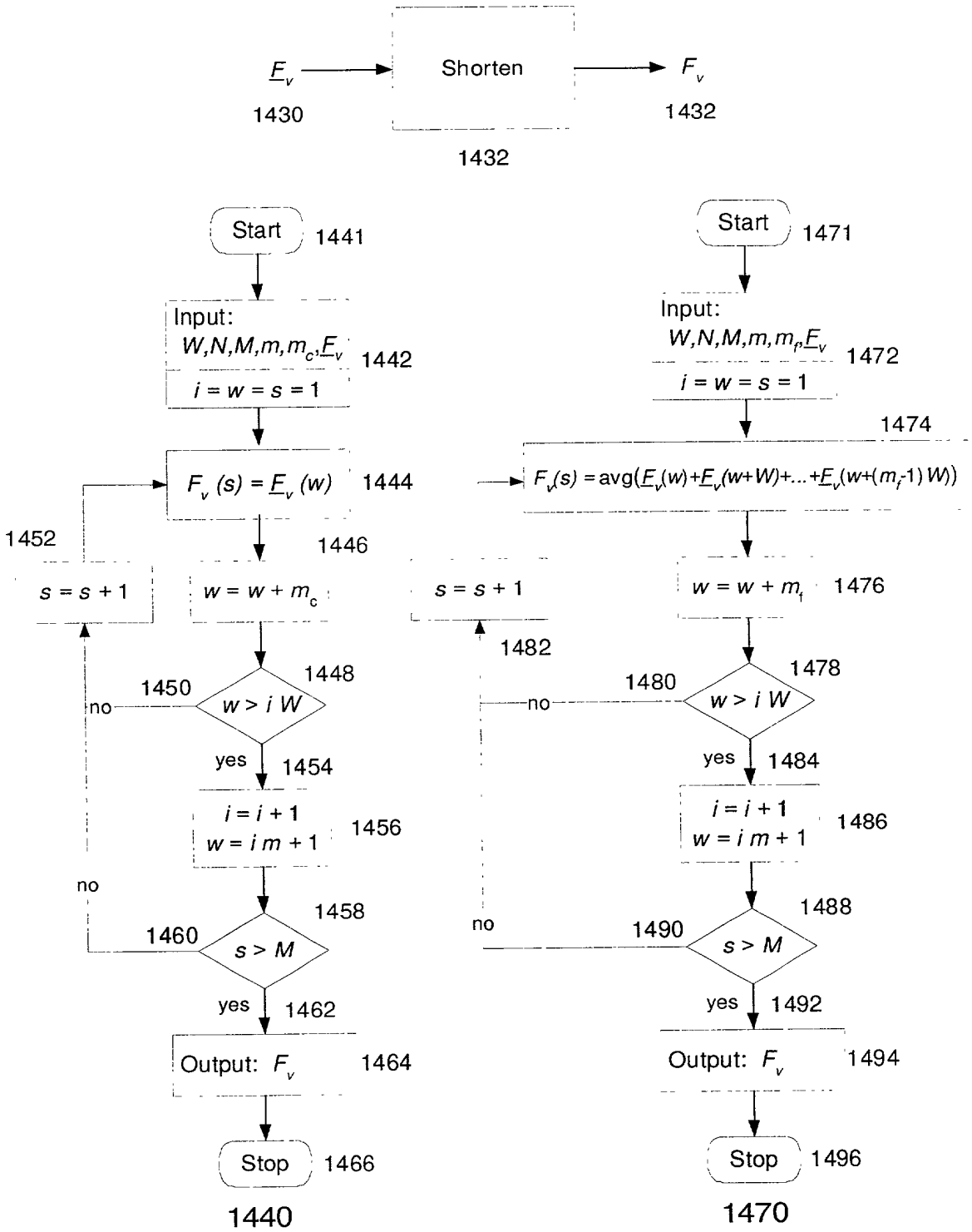

$\underline{F}_v$
1430
Shorten
1432
F_v
1432
Start
1441
Input:
$W,N,M,m,m_c,\underline{F}_v$
1442
$i = w = s = 1$
$F_v(s) = \underline{F}_v(w)$
1444
1452
1446
$s = s + 1$
$w = w + m_c$
1450
1448
no
$w > i\,W$
yes
1454
$i = i + 1$
$w = i\,m + 1$
1456
no
1458
1460
$s > M$
yes
1462
Output: F_v
1464
Stop
1466
1440
Start
1471
Input:
$W,N,M,m,m_f,\underline{F}_v$
1472
$i = w = s = 1$
1474
$F_v(s) = \text{avg}(\underline{F}_v(w)+\underline{F}_v(w+W)+\ldots+\underline{F}_v(w+(m_f-1)\,W))$
$s = s + 1$
1482
$w = w + m_f$
1476
1480
1478
no
$w > i\,W$
yes
1484
$i = i + 1$
$w = i\,m + 1$
1486
no
1488
1490
$s > M$
yes
1492
Output: F_v
1494
Stop
1496
1470

Figure 14B

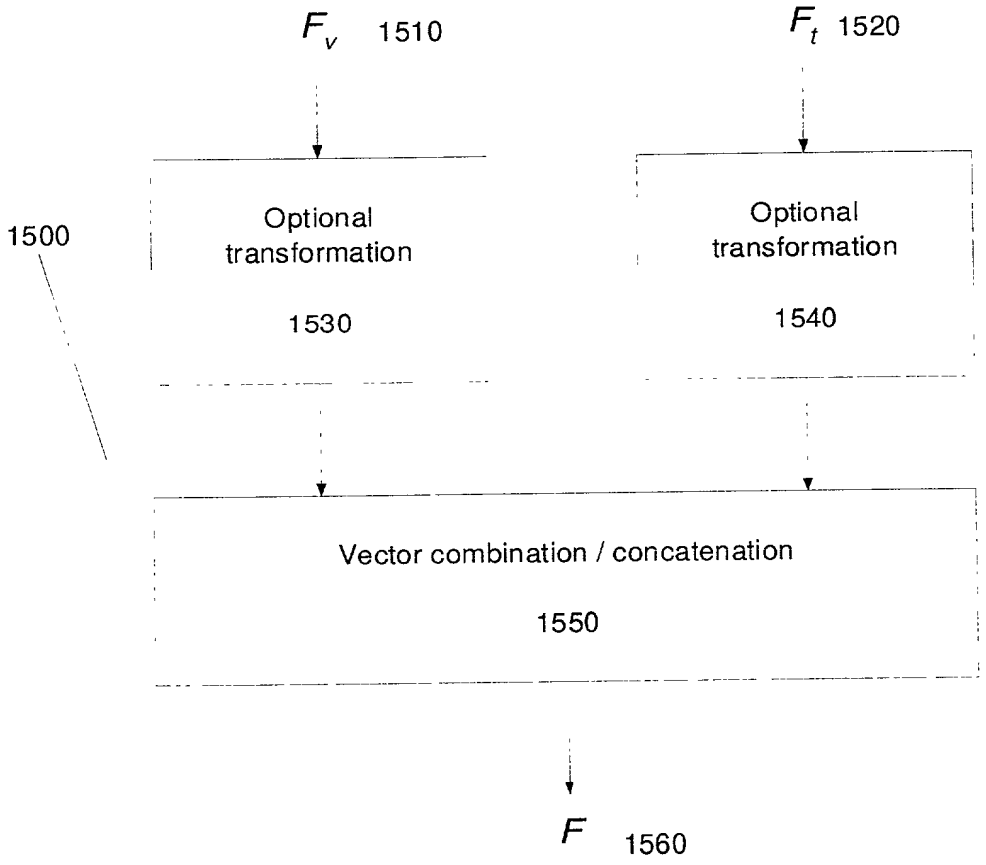

F_v 1510
F_t 1520
1500
Optional transformation
1530
Optional transformation
1540
Vector combination / concatenation
1550
F 1560

Figure 15

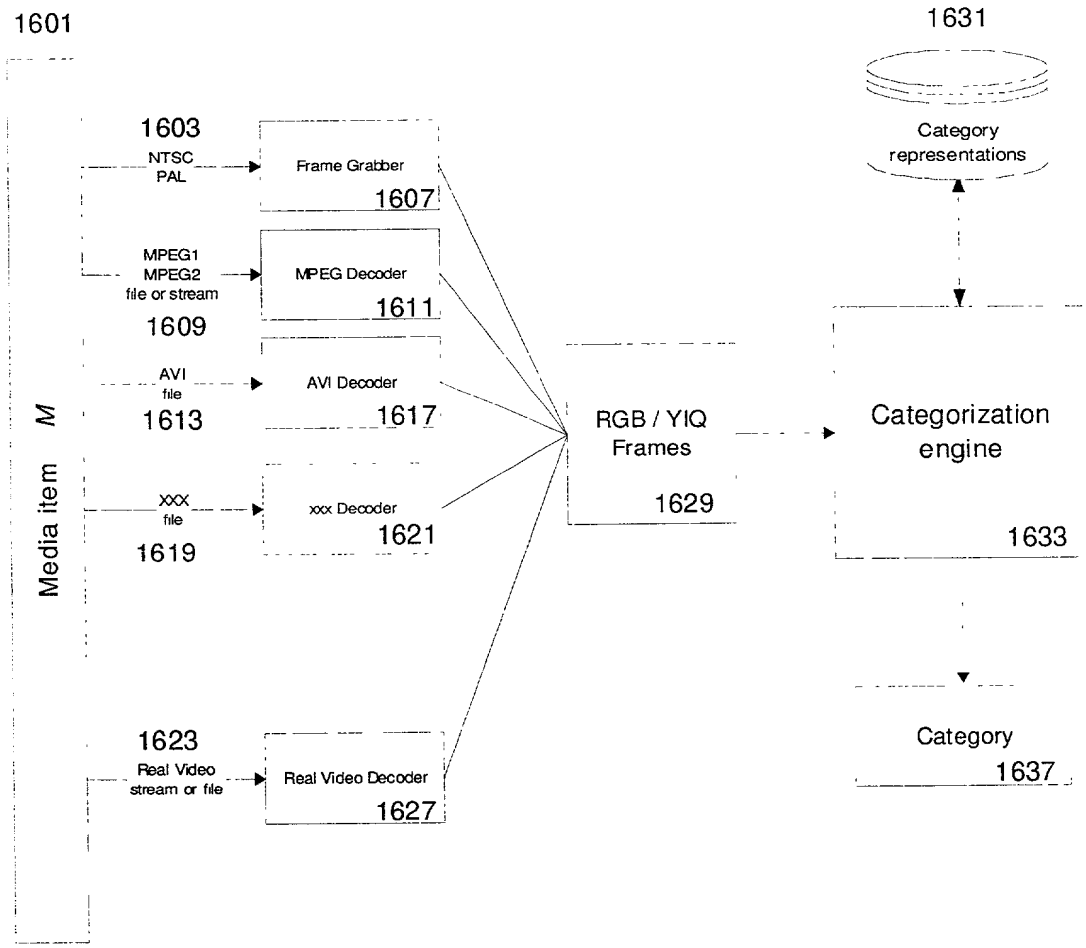
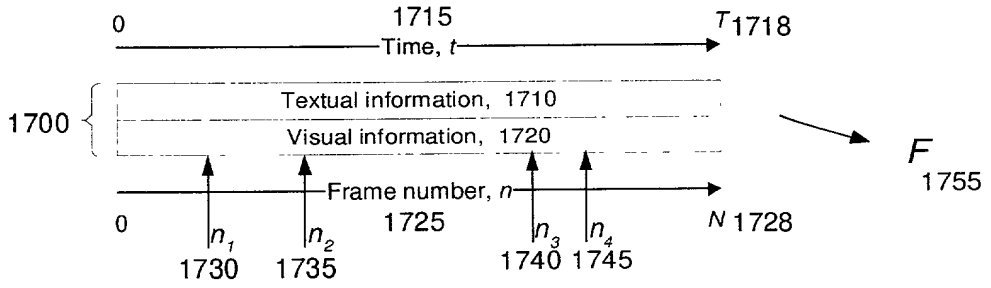
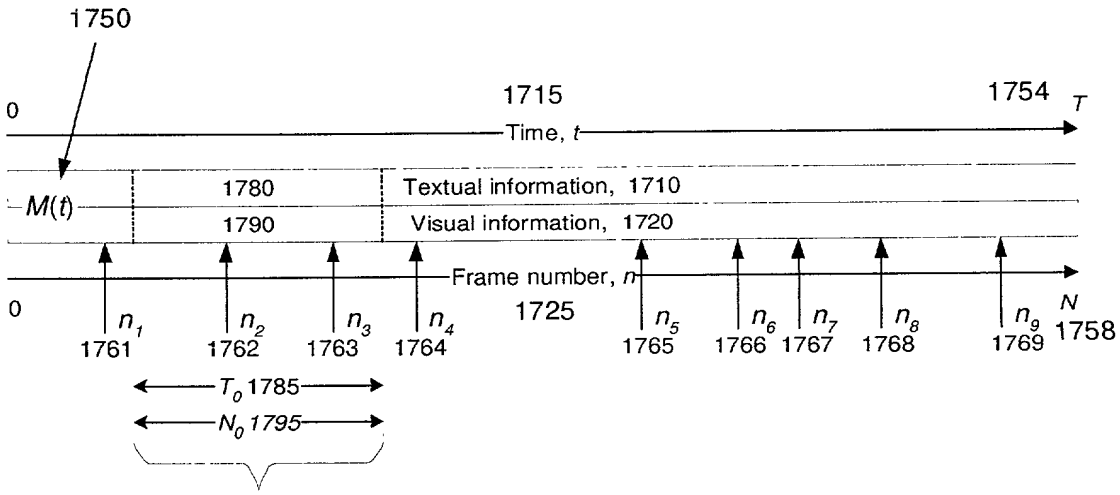

1601
Media item M
1603
NTSC
PAL
Frame Grabber
1607
MPEG1
MPEG2
file or stream
MPEG Decoder
1611
1609
AVI
file
AVI Decoder
1617
1613
XXX
file
xxx Decoder
1621
1619
1623
Real Video
stream or file
Real Video Decoder
1627
RGB / YIQ
Frames
1629
1631
Category
representations
Categorization
engine
1633
Category
1637

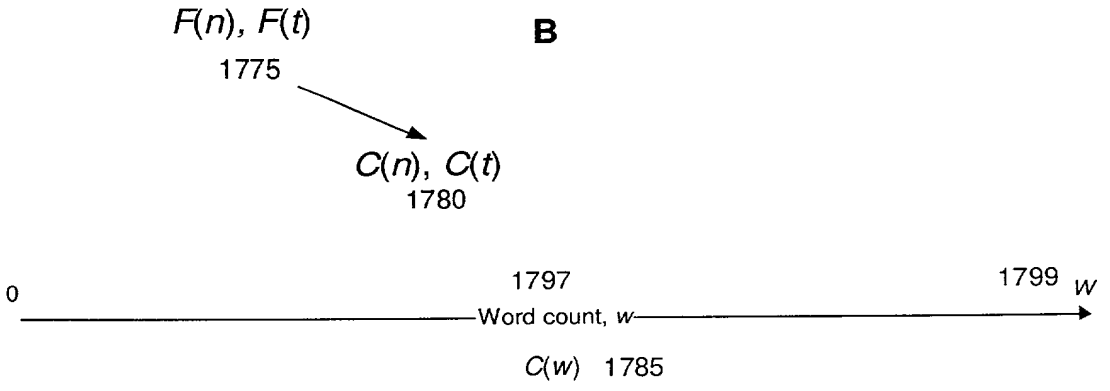
Figure 16



A



B



C

Figure 17

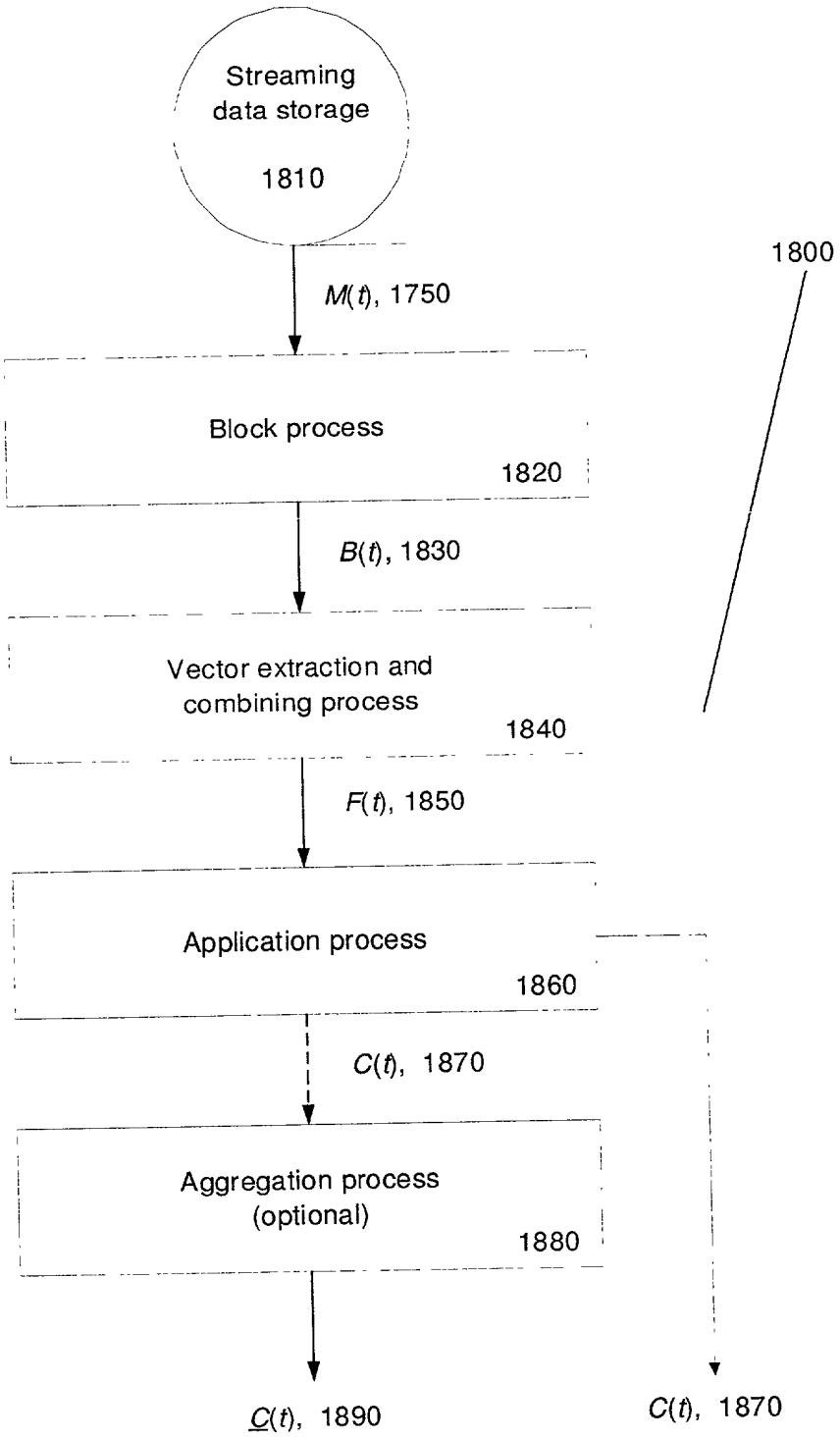
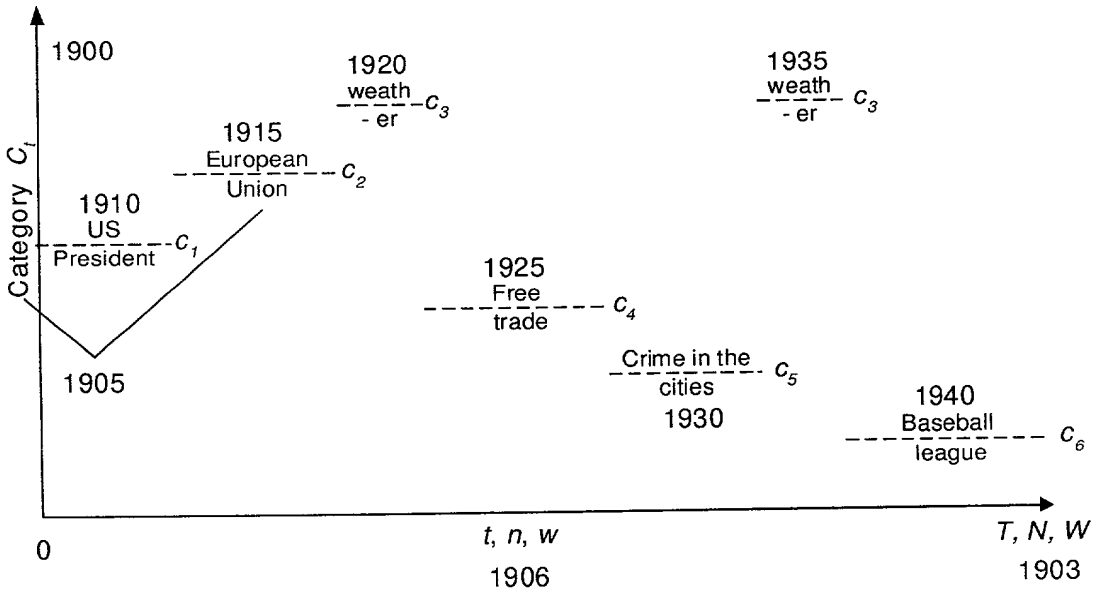
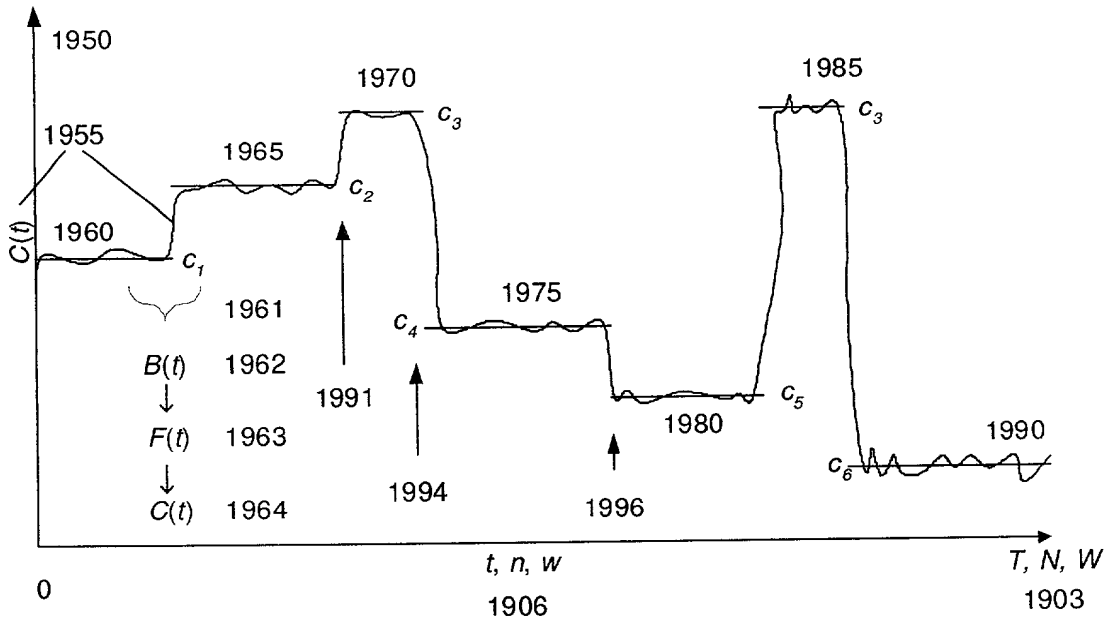

Streaming
data storage
1810
1800
M(t), 1750
Block process
1820
B(t), 1830
Vector extraction and
combining process
1840
F(t), 1850
Application process
1860
C(t), 1870
Aggregation process
(optional)
1880
C(t), 1890
C(t), 1870

Figure 18

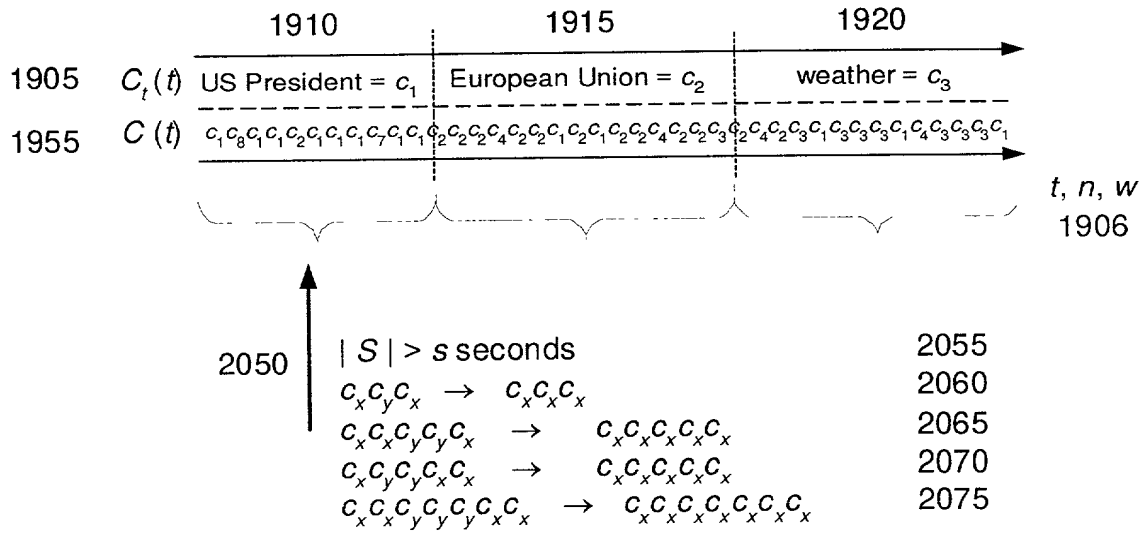

1900
1920
weath
- er
c_3
1935
weath
- er
c_3
1915
European
Union
c_2
Category C_i
1910
US
President
c_1
1925
Free
trade
c_4
Crime in the
cities
c_5
1930
1940
Baseball
league
c_6
1905
0
t, n, w
1906
T, N, W
1903

A


1950
1970
c_3
1985
c_3
1955
1965
c_2
$C(t)$
1960
c_1
1961
1975
c_4
$B(t)$
1962
1991
1980
c_5
1990
$F(t)$
1963
1994
1996
c_6
$C(t)$
1964
0
t, n, w
1906
T, N, W
1903

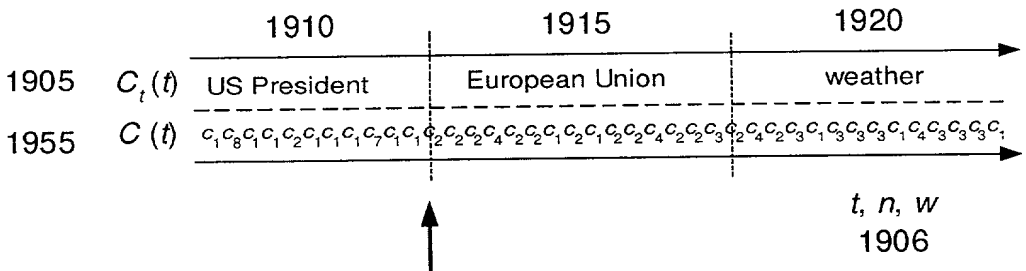
B

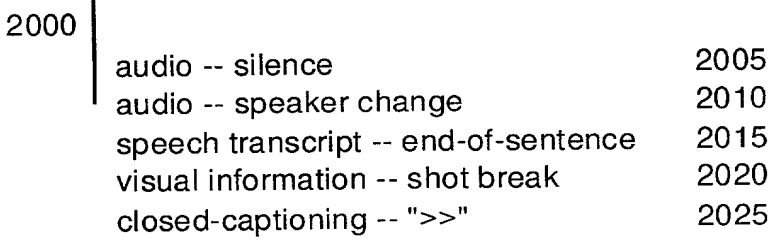
Figure 19


1910
1915
1920
1905
$C_t(t)$
US President = c_1
European Union = c_2
weather = c_3
1955
$C(t)$
t, n, w
1906

2050
| S | > s seconds
2055
$c_x c_y c_x \rightarrow c_x c_x c_x$
2060
$c_x c_x c_y c_y c_x \rightarrow c_x c_x c_x c_x c_x$
2065
$c_x c_y c_y c_x c_x \rightarrow c_x c_x c_x c_x c_x$
2070
$c_x c_x c_y c_y c_y c_x c_x \rightarrow c_x c_x c_x c_x c_x c_x c_x$
2075

A


1910
1915
1920
1905
$C_t(t)$
US President
European Union
weather
1955
$C(t)$
t, n, w
1906


2000
audio -- silence
2005
audio -- speaker change
2010
speech transcript -- end-of-sentence
2015
visual information -- shot break
2020
closed-captioning -- ">>"
2025

B

Figure 20

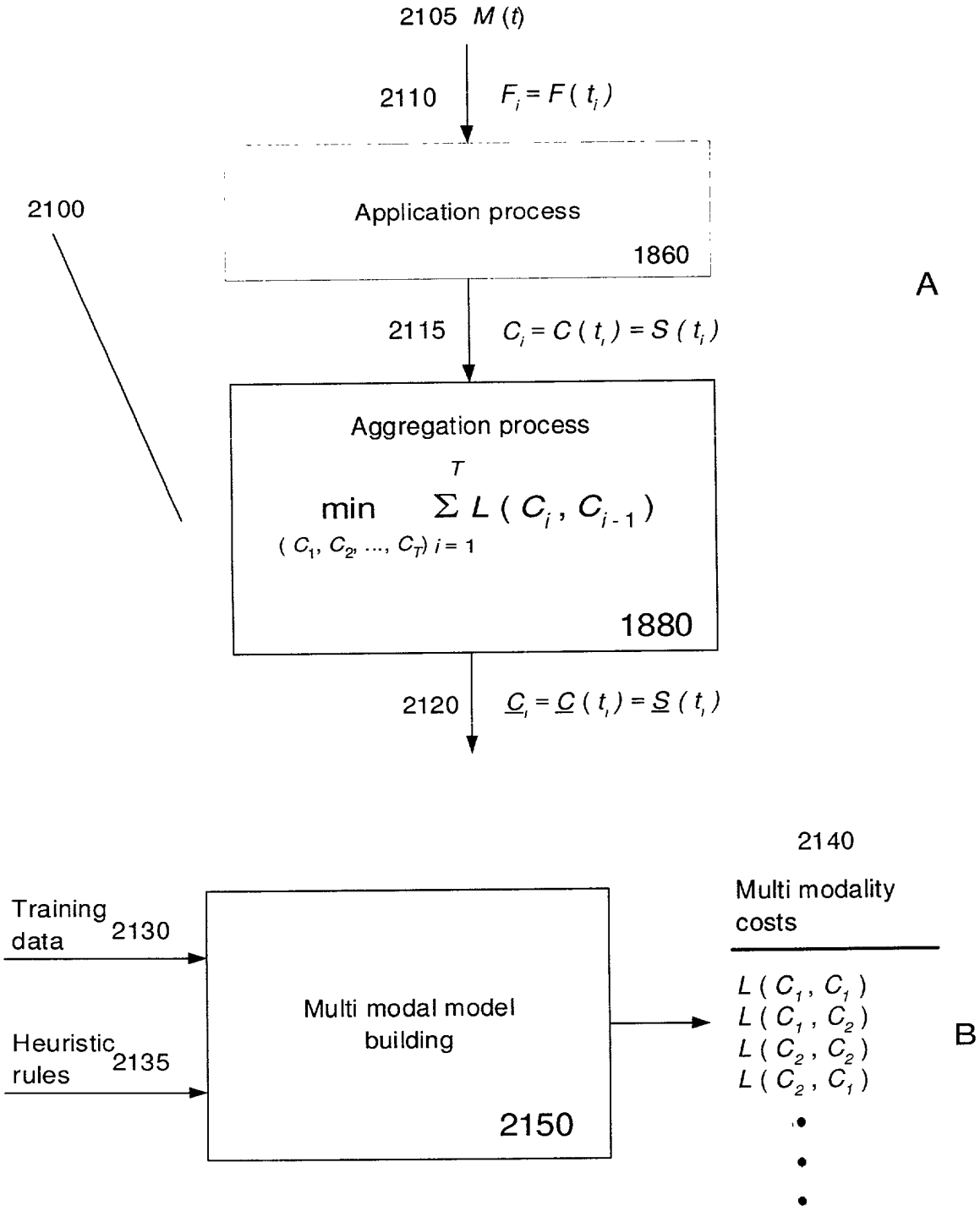

2105 M (t)
2110
$F_i = F(t_i)$
2100
Application process
1860
A
2115
$C_i = C(t_i) = S(t_i)$
Aggregation process
$\min_{(C_1, C_2, ..., C_T)} \sum_{i=1}^{T} L(C_i, C_{i-1})$
1880
2120
$\underline{C}_i = \underline{C}(t_i) = \underline{S}(t_i)$
2140
Multi modality costs
Training data 2130
Heuristic rules 2135
Multi modal model building
2150
$L(C_1, C_1)$
$L(C_1, C_2)$
$L(C_2, C_2)$
$L(C_2, C_1)$
B

Figure 21

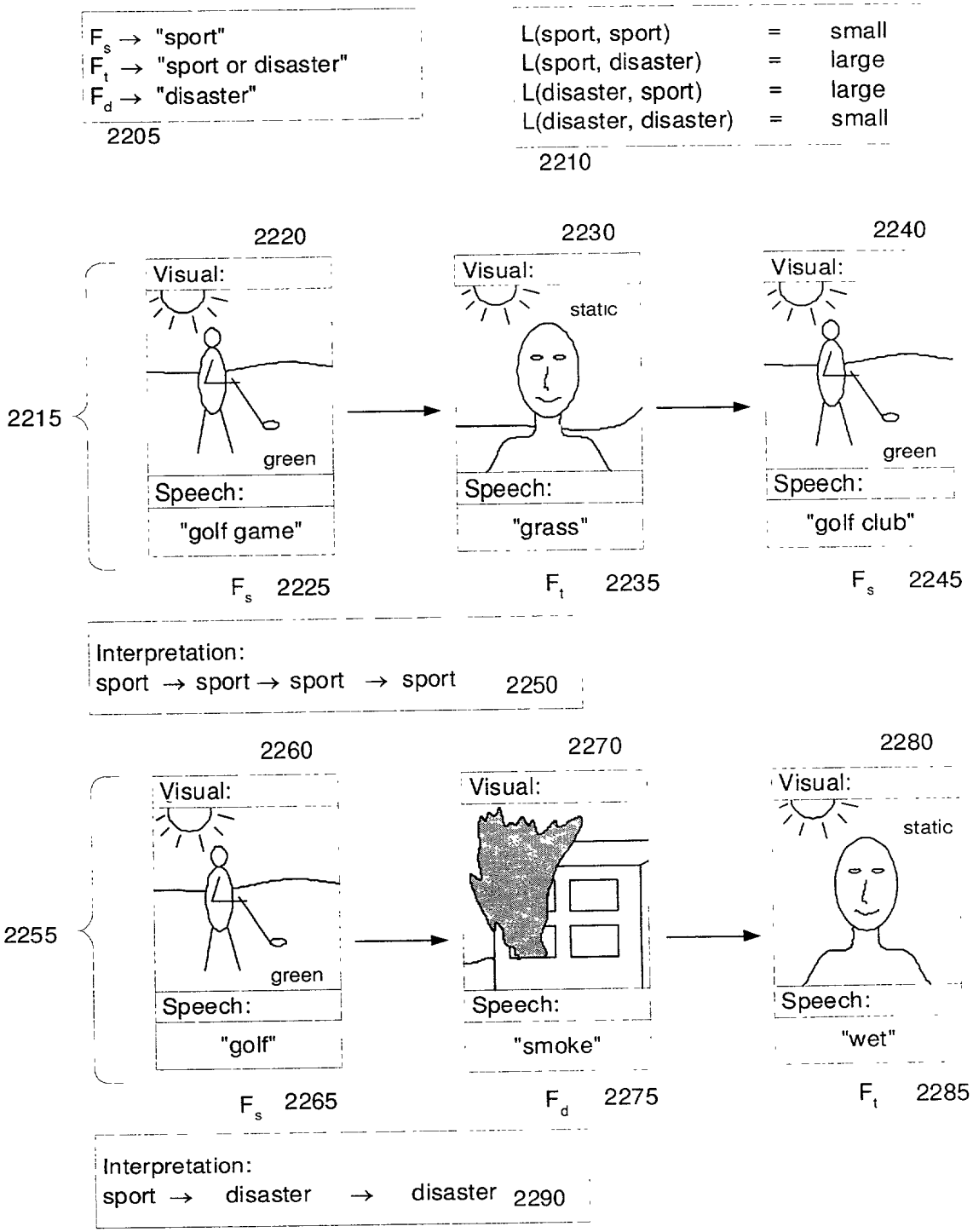

F_s → "sport"
F_t → "sport or disaster"
F_d → "disaster"
2205
L(sport, sport) = small
L(sport, disaster) = large
L(disaster, sport) = large
L(disaster, disaster) = small
2210
2220
2230
2240
2215
Visual:
green
Speech:
"golf game"
Visual:
static
Speech:
"grass"
Visual:
green
Speech:
"golf club"
F_s 2225
F_t 2235
F_s 2245
Interpretation:
sport → sport → sport → sport 2250
2260
2270
2280
2255
Visual:
green
Speech:
"golf"
Visual:
Speech:
"smoke"
Visual:
static
Speech:
"wet"
F_s 2265
F_d 2275
F_t 2285
Interpretation:
sport → disaster → disaster 2290

Figure 22